DOCUMENT RESUME

ED 390 269 FL 023 461

AUTHOR Templin, Stephen A.
TITLE Reliable & Valid Testing of Productive Vocabulary: Speaking Vocabulary Test (SVT).
PUB DATE 95
NOTE 23p.; Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (29th, Long Beach, CA, March 26-April 1, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; *English (Second Language); Higher Education; *Language Tests; Limited English Speaking; Native Speakers; Second Language Instruction; *Second Languages; *Speech Skills; Statistical Analysis; Test Reliability; Test Validity; *Vocabulary; Vocabulary Development

ABSTRACT

This study investigated the testing of speaking vocabulary in English as a Second Language (ESL) at a university in Hawaii. A Speaking Vocabulary Test (SVT) was developed and piloted with college students. Test-takers (n=37) were divided into three groups: native English-speaking freshmen and sophomores; non-native English-speaking freshmen, sophomores, juniors, and seniors; and non-native English-speaking students enrolled in an intensive English program preparatory to mainstream university classes. Results indicate the test to be reliable and valid: students' scores were consistent, showing a high level of correlation; two evaluators' scoring of the same ten random tests showed high correlation; and the three student groups had significantly different scores, ranking in descending order: native; non-native; and non-native language institute. Anecdotal information on student response to the test is also offered. A brief bibliography, the test, and test evaluator guidelines are appended. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TESTING SPEAKING VOCABULARY

ED 390 269

Reliable & Valid Testing of Productive Vocabulary: Speaking Vocabulary Test (SVT)

Stephen A. Templin

Brigham Young University-Hawaii

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Stephen A. Templin
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

Author note: Stephen A. Templin (now at Kanagawa Prefectural Board of Education, Japan Ministry of Education), Teaching English to Speakers of Other Languages. This study was funded in part by a research grant from Brigham Young University-Hawaii. The author wishes to thank Chad Compton and Andrew Allen for their assistance in this research. Part of this study was presented at the 1995 TESOL Convention in Long Beach, California.

Correspondence should be addressed to Stephen A. Templin, Green Valley A-105, 1937 Ohzenji, Asao-ku Kawasaki-shi, Kanagawa-ken 215, JAPAN.

FL023461

BEST COPY AVAILABLE

Abstract

Vocabulary, particularly speaking and writing vocabulary, has been frequently ignored in L2 research. This study examines speaking vocabulary's relationship with productive and receptive vocabulary. Because teachers and students need a reliable and valid speaking vocabulary test, a test for measuring speaking vocabulary has been created: *Speaking Vocabulary Test (SVT)*. The SVT proves to be a reliable and valid test: (1) Students' SVT scores were consistent, showing a significantly high level of correlation, .83. (2) Two evaluators' scoring of the same ten random tests showed a significantly high level of correlation, .97. (3) The SVT proved valid in that three groups of students, native, nonnative, and nonnative English Language Institute (ELI), had scores ranking in descending order: native, nonnative, and nonnative ELI. Also, in an ANOVA, the difference between the scores of the three groups was significant at the <.05 level.

Introduction

It is hard to imagine language without vocabulary. Allen (1983) says that ESL students need good vocabulary to communicate effectively; without good vocabulary, communication breaks down. However, studies in second language (L2) acquisition focus on morphology and syntax, neglecting vocabulary. According to Larsen-Freeman and Long (1993), L2 studies frequently ignore vocabulary.

The little vocabulary research which has been done focuses mostly on receptive (listening and reading) vocabulary rather than productive vocabulary (speaking and writing). It is not enough to be able to read and listen to vocabulary; students must speak and write vocabulary, too. In addition, teachers need a reliable and valid measurement of students' productive vocabulary. In testing productive vocabulary, both speaking and writing are important; however, this study will be limited to testing speaking vocabulary.

Productive and Receptive Vocabulary

Previous vocabulary research has made a distinction between productive and receptive vocabulary. Previous researchers have described productive vocabulary with a variety of terms: active and passive, use and understanding, recall and recognition, real and potential, encoding and decoding, etcetera. For consistency within this study, the term production will refer to vocabulary used in speaking and writing, while reception will represent listening and reading vocabulary.

Rather than being separate entities, research suggests that production and reception are interrelated. Researchers report this to be a process where learners receive vocabulary through listening and reading, then, over time, part of that vocabulary moves gradually toward production via speaking or writing (Faerch, Haastrup, and Phillipson, 1984; McCarthy, 1990; Meara, 1990; and Palmberg, 1990).

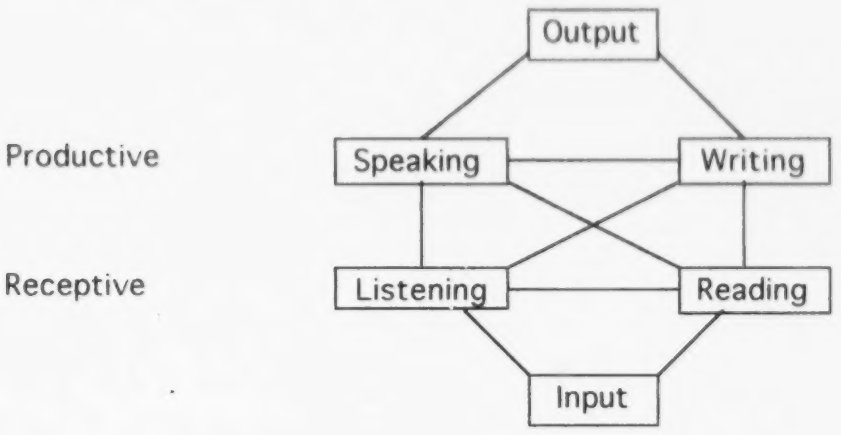
Although production and reception are interdependent, not all receptive vocabulary will be produced. Students, teachers, and researchers generally agree that language learners have less productive than receptive vocabulary. ESL learners may be able to recognize a word they hear or read, but that word may not become part of their speaking or writing. In Melka-Teichroew's (1982) comprehensive survey of productive and receptive vocabulary, she cites other researchers who claim that productive vocabulary in ESL learners is about two times smaller than receptive vocabulary. Allen (1983) estimates the size of productive vocabulary to be as much as 10 times less than receptive vocabulary. Nation (1990) says that (for native English speakers) producing a word is 50 to 100 percent more difficult than understanding it. Interestingly, research in language attrition, language loss, shows that when people forget vocabulary, production of words is lost before reception (Cohen, 1989). In spite of the different estimates of productive verses receptive vocabulary, most agree that language learners can hear or read a word before they speak or write it. Also, the pool of vocabulary items produced via speaking and

writing is less than the pool of vocabulary understood through listening or reading.

The model below may be helpful in summarizing the relationship between productive and receptive vocabulary and the connection between speaking, writing, listening, and reading.

Figure 1

ESL Vocabulary Productive/Receptive Model



Most ESL learners receive vocabulary through listening or reading, first. The new vocabulary enters one or both receptive areas before speaking or writing production. An ESL learner does not usually speak or write a new vocabulary word until they have heard or read it first.

The lines in Figure 1 above show the connections of the four vocabulary areas: speaking, writing, listening, and reading. These lines represent a continuum of movement from one area to another rather than some magical transformation from place to

place. When learners receive vocabulary from input, they can transfer a word from listening to reading or vice versa. For example, if learners acquire a new word through listening, it is possible that they will understand that word when they come across it while reading. Similarly, during a listening act, they might remember a word they acquired through reading.

For a word to become part of productive vocabulary, it moves along the continuum toward production until it can be activated for use in speaking or writing. Vocabulary production in one skill may influence production in another skill. For instance, learners might speak a word that influences them to write the same word. The same is true of the opposite; a written word might influence production in the spoken form. In summary, speaking, writing, listening, and reading are all interconnected: a word is inputted into listening or reading and then can move along a continuum to speaking and writing production for output.

Testing Speaking Vocabulary

Although productive and receptive vocabulary skills are both important, Hughes (1989) writes that productive vocabulary testing is rare. He says, "Information on receptive ability is regarded as sufficient" (149). While measurements of receptive vocabulary give an indirect indication of what productive vocabularies *might* be (an unknown number of vocabulary less than the receptive number), they do not give a direct measurement of what students' productive vocabulary actually is.

Because teachers do not test productive vocabulary, students

and teachers are left in uncertainty. Students with a productive vocabulary less than average for their level should know that they need to improve. Students who rely on overusing certain words or repeatedly do not use the appropriate words could benefit from feedback that measures their productive vocabulary level. Another positive aspect of this feedback for students is that they can feel a sense of accomplishment for the progress they have made in increasing their vocabulary production, motivating them to continue their progress.

For teachers, productive vocabulary testing is beneficial in two ways. First, it helps them find out if their students are at a reasonable level for the goals of the class. Second, testing would be useful to gauge the effectiveness of teaching methods and materials for helping students improve their productive vocabulary.

There have only been a few measurements of productive vocabulary, particularly speaking, developed in the past. One type of measurement, used by Arnaud (1984) and Laufer (1991), tested vocabulary richness in writing; however, in creating a variation of this measurement for speaking, testing vocabulary richness was considered too time consuming with ambiguous results. Next, a modified version of Corson's (1989) two-word test was created and employed on a trial run. This modified version was named the Speaking Vocabulary Test (SVT) and considered feasible for administration to a larger population of subjects.

Reliability & Validity

In order for a test such as the SVT to be effective, it must be reliable and valid. For reliability, students test scores should be consistent, not failing one moment and passing the next. One way of checking this consistency is by using the split-half method: dividing students' scores into two tests, the odd numbered responses being one test and the even numbered responses being the other test. These two tests are then compared to see if there is a significant correlation between them. Highly significant correlation shows that the test is reliable.

Another way of checking for reliability, called interrater reliability, is to see if the test evaluators are consistent in their scoring. Two evaluators' scores of the same tests are compared. A reliable test has a highly significant correlation between the scores given by both evaluators.

Besides reliability, validity is also important in a test. A test must measure what is says it will measure to be valid. For this study, university students were divided into three groups: (1) native English speaking freshmen and sophomores, (2) nonnative English speaking freshmen, sophomores, juniors, and seniors, and (3) nonnative English speaking students enrolled in the English Language Institute (ELI), preparatory to mainstream university classes. If the SVT is valid, it is expected that when the scores of these three groups are compared, native English speakers will score the highest, nonnative English speakers will be second, and nonnative ELI students will have the lowest score.

Three hypotheses were formulated: (1) The SVT will prove

reliable in that the students' test scores will show a significantly high level of correlation. (2) The SVT will prove reliable in that two evaluators' scoring of the same ten random tests will show a significantly high level of correlation. (3) The SVT will prove valid in that three groups of students, native, nonnative, and nonnative ELI, will have scores in descending order: native, nonnative, and nonnative ELI. In addition, the difference between the scores of these three groups will be significant at the <.05 level.

METHOD

Subjects

At a university in Hawaii, 37 students were selected. These students were (1) native English speaking freshmen and sophomores, (2) nonnative English speaking freshmen, sophomores, juniors, and seniors, and (3) nonnative English speaking students enrolled in the English Language Institute (ELI), preparatory to mainstream university classes. The native English speakers came from a variety of ethnic backgrounds. The nonnative English speakers came from various countries in Asia, the Pacific Islands, South America, and Europe.

Materials

Using Nation's (1990) list of university words, 60 words were randomly selected and formed into 30 pairs (see Appendix B). Each pair was printed in bold lettering on one 3x5 note card for a total of 30 cards. Tape recorders and cassettes were used to record student responses. To make sure students did not go over a

12-minute time limit the evaluators used the stopwatch function on their watches.

Instructions for the SVT with examples of acceptable responses (Appendix A) were made. Evaluator guidelines (Appendix C) and a list of the words used in the administration of SVT (Appendix B) were created for the evaluators, also.

Procedures

Two proctors administered the tests, separately, to one student at a time. First, a student taking the test filled out a demographic sheet. The proctor documented the name of the student on the cassette to be used in recording. Then, the proctor read the SVT instructions, answering any questions the student asked.

When the student was ready, the proctor pressed record on the tape recorder, started the timer, and showed the first combination of words to the student (the order of pairs given was the same as in Appendix B). The proctor only showed one card at a time. When a student gave a sentence or passed, the proctor removed the card and showed the student the next combination. This process continued until each student finished all thirty combinations or 12 minutes expired, whichever came first. After recording all responses, the responses were transcribed, the names of the students were removed, and the transcribed responses were evaluated. Later, two evaluators, Evaluator 1 and Evaluator 2, evaluated a random selection of ten tests to establish interrater reliability. The evaluators were given the instructions for the test (Appendix A), the list of word pairs tested (Appendix B), and

the evaluator guidelines (Appendix C). Evaluator 2, unfamiliar with the test, received about five minutes of training before examining the tests. Although both evaluators referred to dictionaries, what dictionary to use was not specified.

RESULTS

The three hypotheses were shown to be true: (1) The SVT proved reliable in that the students' test scores showed a significantly high level of correlation, .83. (2) The SVT proved reliable in that the two evaluators' scoring of the same ten random tests showed a significantly high level of correlation, .97. (3) The SVT proved valid in that three groups of students, native, nonnative, and nonnative ELI, had scores in descending order: native, nonnative, and nonnative ELI. Also, the difference between the scores of the three groups was significant at the $<.05$ level.

Reliability was determined by analyzing the consistency in the students' scores. First, using the split-half method, each of the 37 students' SVT scores was divided into odd and even numbered responses, shown in Table 1.

Table 1

Reliability: Split-half Scores

Overall Scores	Odd Scores	Even Scores
26	14	12
24	13	11
12	07	05
10	05	05
21	11	10
10	08	02
15	08	07

10	04	06
24	13	11
26	14	12
11	06	05
14	09	05
01	01	00
22	11	11
19	10	09
23	14	09
08	06	02
14	08	06
14	09	05
25	12	13
21	13	08
04	01	03
22	12	10
01	01	00
22	12	10
08	04	04
10	04	06
24	13	11
08	06	02
10	06	04
05	04	01
16	10	06
30	15	15
12	03	09
23	14	09
05	02	03
07	03	04

The odd and even scores from above were correlated in the matrix below.

Table 2

Reliability: Correlation of Split-half SVT Scores

	Test 1	Test 2
Test 1	1.00	.83*
Test 2	.83*	1.00

*Significant

The students' scores were consistent with a significantly high correlation of .83.

In addition to examining students' consistency, interrater reliability was determined by correlating the scores given by two evaluators to ten students as shown in Table 3.

Table 3

Validity: Interrater Reliability of Scoring by Evaluators 1 & 2

Scores Given by Evaluator 1	Scores Given by Evaluator 2
26	24
24	23
12	10
10	04
21	20
10	07
15	11
10	11
24	22
26	25

	Evaluator 1	Evaluator 2
Evaluator 1	1.00	.97*
Evaluator 2	.97*	1.00

*Significant

The correlation of the two evaluators' scoring has a significantly high correlation, .97.

After analyzing reliability, validity was studied. Individual students' test scores, the highest possible score being 30 and the lowest score being 0, were divided into three groups: native, nonnative, and nonnative ELI (See Table 4).

Table 4

Validity: SVT Scores of Native, Nonnative, & Nonnative ELI Students

	Native	Nonnative	Nonnative ELI
Students'	26	21	12
SVT Scores	24	07	10

(1-30 Points)			
	24	23	10
	26	08	15
	14	21	10
	22	05	11
	19	23	01
	14		14
	25		04
	22		01
	22		08
	24		10
	16		08
	30		10
			05
			12

An ANOVA was used to compare the scores for native, nonnative, and nonnative ELI students. The results are reported in Table 5.

Table 5

Validity: One-way ANOVA

	DF	Sum of Squares	Mean Squares	F	p
Level	2	1299.7	649.9	23.01	0.00
Error	34	960.2	28.2		
Total	36	2259.9			

	N	Mean	SD
Native	14	22	4.72
Nonnative	7	15.42	8.28
Nonnative ELI	16	8.81	4.15

The mean for the native English speakers was higher than both nonnative levels. Also, the nonnative students in mainstream classes had a higher average than the nonnative ELI students.

To see if these differences were significant at the <.05 level, a Tukey comparison was computed for each level, shown in Table 6.

Table 6

Validity: Tukey Comparison of Levels for Significant Differences

Level	Intervals	<.05 Significance

Native & Nonnative	-12.607 to -.536	Significant
Nonnative & Nonnative ELI	-12.525 to -.707	Significant
Native & Nonnative ELI	-17.959 to -8.416	Significant

Intervals not containing 0 = Significant

Because the confidence intervals for the differences at the <.05 level do not contain 0 for any of the levels, all of the levels are significantly different.

DISCUSSION/RECOMMENDATIONS

The Speaking Vocabulary Test (SVT) is a reliable and valid test of ESL students' speaking vocabulary. During the administration of SVT, several students asked what score is appropriate for their level. Others wanted to take the test during the next semester to see how much they improve. Some said they felt that testing like this would motivate them to try harder in speaking new vocabulary. A couple showed surprise at how low their scores were. Overall, most students expressed that the SVT was testing their speaking vocabulary, not something else.

Many students said that they knew one of the words in the question pair, but not both. Some word combinations may seem too difficult, but the students who had a greater mastery of vocabulary in this study could produce sentences that showed the meaning of both words. The more difficult combinations served as discriminators between students with higher and lower productive vocabulary levels.

An advantage of using the two-word method, as opposed to only one word, is that the combination of word pairs possible is so numerous that students will not be able to get by with rote

answers of memorized sentences. They have to know how to show the meaning of the vocabulary words given. Although Nation's university word list was used for this study, teachers can use any list that fits the specific needs of their students.

One problem with this test is the time involved in administering, transcribing, and grading the tests. The proctors tested the students individually to have greater control over intervening variables that might affect the research and to allow the proctors better observation of the testing process. Also, the time limit of 12 minutes could be reduced. The use of a language lab should cut down the time needed for administering the tests. Students completed the test in an average of 8 minutes, so the time allowed to complete the test could be shortened.

In this study, because of scheduling constraints, only native English speaking freshmen and sophomores took the SVT. It would be interesting to compare how native English speaking juniors and seniors score on the SVT.

REFERENCES

Allen, V. F. (1983). Techniques in teaching vocabulary. New York: Oxford UP.

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. Papers from the International Symposium on Language Testing, 7. Colchester, England. (ERIC Document Reproduction Service No. ED 275 164)

Cohen, A. D. (1989). Attrition in the productive lexicon of two Portuguese third language speakers. Studies in Second Language Acquisition, 11, 135-49.

Corson, D. (1989). Adolescent lexical differences in Australia and England by social group. Journal of Educational Research, 82, 146-157.

Faerch, C., Haastrup, K., and Phillipson, R. (1984). Learner Language and Language Learning. Clevedon: Multilingual Matters.

Hughes, Arthur. (1992). Testing for Language Teachers. Cambridge: Cambridge UP.

Larsen-Freeman, D. and Long, M. H. (1993). An Introduction to Second Language Research. New York: Longman.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. The Modern Language Journal, 75, 440-448.

McCarthy, M. (1990). Vocabulary. Oxford: Oxford UP.

Meara, P. (1990). A note on passive vocabulary. Second Language Research, 6, 150-54.

Melka-Teichroew, F. J. (1982). Receptive vs. productive vocabulary: A survey. Interlanguage Studies Bulletin Utrecht, 6, 5-33.

Nation, P. (1990). Teaching and Learning Vocabulary New York: Newbury House.

Palmberg, R. (1990). Improving foreign-language learners' vocabulary skills. RELC Journal, 21, 1-10.

APPENDIX A

Speaking Vocabulary Test (Read/Speak)

Directions:

-You will have 12 minutes to complete 30 problems. Each problem will list two words on an index card.
Example: *type/computer.*

-Say a complete sentence using both words.
Example: A person can *type* on a *computer* .

-If you know that you cannot create a sentence for one pair, pass and move on to the next. (If you pass, you do not get another chance at problems you pass.)

-Try to complete as many as you can, as accurately as possible.

-You can say the two words in any order.
Example: A person who uses a *computer* knows how to *type.*

-Also, you may say any form of the word.
Example: *Typing* on *computers* is not difficult.

-Although some words have more than one meaning, say a sentence that uses the meaning you want.
Example: There is more than one *type* of *computer* for sale.

-Do not use proper nouns or adjectives: I *typed* on a new keyboard at *Computer* Land. (*Computer* is part of a store's name.)

APPENDIX B*

1. withdraw/bubble
2. devise/vocabulary
3. ignore/violate
4. volume/impose
5. harbor/prosper
6. manifest/volt
7. monarch/distribute
8. explicit/mobile
9. cancel/converse
10. portion/series
11. contemplate/capture
12. adjust/code
13. precede/adequate
14. implement/administer
15. drama/abnormal
16. vital/vague
17. define/generate
18. export/ratio
19. internal/interrelate
20. acquire/apparatus
21. rely/assure
22. conflict/factor
23. psychology/assign
24. synthetic/saturate
25. cater/consist
26. convene/equate
27. minor/incident
28. revive/congress
29. tone/authorize
30. supplement/rotate

*These combinations were randomly selected from Nation's (1990) University Word List (p. 235-239).

APPENDIX C

Speaking Vocabulary Test Evaluator

Students are expected to produce a sentence which indicates a correct definition of the given words.

A. Review test and problem items.

B. Evaluate student responses.

1. Students make *one sentence using both words.*

   a. More than 1 sentence is wrong.

   b. A sentence using only one word of the pair is wrong.

   c. Using the wrong word is wrong. (Different tenses or forms of the correct word are correct.)

   d. Pass is wrong.

2. The following responses are not acceptable:

   a. Indicating the definition of only one word of the pair is wrong.

   b. Indicating an incorrect definition of either word is wrong.

   c. Sentences which *ask* for the definition of either word are wrong.
   (Example: *What is the meaning of devise and vocabulary?*)

   d. Sentences which avoid indicating the definition of either word are wrong.
   (Example: *Devise is a vocabulary word.* This sentence avoids indicating the definition of *devise*.

3. These variations are acceptable:

   a. Incorrect grammar is acceptable.
   (Example: *Teacher device vocabulary test for students.*)

   b. Sentences which are somewhat vague, yet indicate the correct definition, are acceptable.
   (Example: *My psychology teacher gave me an assignment. Psychology* is a field of study, and an *assignment* is usually given to students by teachers.)

c. Nonstandard usage, which still indicates the definition, is acceptable.
(Example: *Your infant is growing like a grass*. Although *a weed* would be standard, rather than *a grass*, this sentence still indicates the definition of both words.)

4. Scoring student responses.

a. Each acceptable sentence is given 1 point.

b. Each wrong sentence is given 0 points.

c. The maximum possible is 30/30.